1. Shared task and dataset description

The dataset used in this study is derived from the MIMIC-IV (Medical Information Mart for Intensive Care) database [1], a publicly accessible collection of de-identified clinical data from the Beth Israel Deaconess Medical Center and hosted on PhysioNet [2]. It was constructed for the 2024 BioNLP Workshop shared task "DischargeMe!"[1], which focused on generating two key sections of hospital discharge summaries: the Brief Hospital Course (BHC) and Discharge Instructions (DI) [3]. The baseline overall score reported for the task was 0.126. A total of 122 participants registered, and 13 teams appeared on the final leaderboard on Codabench[2], which reports evaluation results on the final test set.

The dataset is publicly available via PhysioNet[3] and was constructed from the MIMIC-IV submodules MIMIC-IV-Note and MIMIC-IV-ED. It includes discharge summaries from 109,168 Emergency Department (ED) visits, enriched with structured data (e.g., demographics, diagnoses, medications, laboratory results, vital signs, and ICD information) and unstructured clinical text (e.g., discharge notes, radiology reports, chief complaints). A unique identifier links records across tables, allowing multiple entries such as radiology reports to correspond to a single patient admission. The target sections (BHC and DI) are provided separately in a dedicated CSV file. Because these sections also appear within the original discharge text, they are removed from the input using the regular expression patterns provided by the task organizers to prevent data leakage during modeling.

The complete dataset is divided into four splits: training (68,785 samples), validation (14,719 samples), Phase I testing (14,702 samples), and Phase II testing (10,962 samples). During development, access to the two test splits was prohibited; Phase II served as the final evaluation set for leaderboard ranking. The organizers released Python code to reproduce the official splits via a Google Colab notebook[4], which requires a valid PhysioNet account and supports both local and cloud execution.

We obtained the 250 unique identifiers corresponding to the Phase II evaluation subset used for final scoring[5].

	Total samples	Mean tokens	Median tokens	Standard deviation	95th percentile
Train	64655	3874.6	3679	1391.15	6567
Valid	13865	3873.24	3674	1407.42	6587
Test 1	13869	3885.9	3697	1400	6593
Test 2	250	2660.52	2520	1005.68	4433.2

Table 1: Data statistics of the prompt, with instruction, input texts (radiology and discharge) and response

	Total samples	Mean tokens	Median tokens	Standard deviation	95th percentile
Train	68711	1819.36	1511	1091.69	3884
Valid	14702	1818.29	1503	1096.21	3866
Test 1	14691	1822.25	1502	1095.13	3900
Test 2	250	1076.2	864.5	710.67	2412

Table 2: Data statistics of the truncated prompt, with instruction, input texts (truncated discharge) and response

2. Statistics of prompts and different fields of input data

In the shared task, four data splits are provided: training, validation, Test Phase 1, and Test Phase 2. Each split contains six CSV files. In addition to structured fields, the data include several free-text fields corresponding to radiology reports and discharge texts, as well as the target texts, BHC and DI. A unique "hadm_id" is used to link records across these files. While each "hadm_id" is associated with a single discharge text and corresponding BHC and DI texts, multiple radiology reports may exist for the same admission. For our experiments, all radiology reports associated with a given "hadm_id" are concatenated and used as input to the model.

Table 3 reports token-length statistics for the raw text fields, computed using the GPT-2 tokenizer. Specifically, we report the average token length, standard deviation (SD), and the minimum and maximum token lengths.

[1] https://stanford-aimi.github.io/discharge-me/

[2] https://www.codabench.org/competitions/2008/#/results-tab

[3] https://physionet.org/content/discharge-me/1.3/

[4] https://colab.research.google.com/drive/1yW-29KcDYoswMrqwjEMO6l6i2Ll3p5BX?usp=sharing

[5] The unique identifier list was obtained from the organizers after the competition and is available in our GitHub repository.

	Radiology text				Discharge text				BHC				DI			
	Average	**SD**	**Min**	**Max**	**Average**	**SD**	**Min**	**Max**	**Average**	**SD**	**Min**	**Max**	**Average**	**SD**	**Min**	**Max**
Train	1354.55	1754.17	29	44644	2981.77	1297.09	223	17259	576.29	424.48	11	6259	309.81	251.63	10	8184
Valid	1341.29	1719.91	29	44682	2984.52	1313.95	351	14417	571.92	426.47	15	6728	312.13	257.83	13	8664
Test Phase1	1338.85	1709	46	34020	2983.23	1293.96	239	16775	576.09	424.86	13	3930	309.71	250.82	14	5074
Test Phase 2	1339.59	1709.98	27	27296	2972.03	1282.22	421	12508	573.51	429.92	12	4309	310.42	254.18	12	4700

Table 3: Average, standard deviation, minimum, and maximum token counts for the training, validation, Test Phase 1, and Test Phase 2 datasets across the four text fields.

Tables 1 and 2 present token-length statistics for the constructed prompts under the full-input and truncated-input strategies, respectively. For each setting, we report the mean, median, standard deviation, and the 95th percentile of prompt lengths. As expected, the truncated-input strategy substantially reduces prompt length, with the mean prompt size being less than half that of the full-input strategy. Since the organizers selected only 250 documents from the Phase 2 test dataset, we also report descriptive statistics computed exclusively on this subset.

In the original dataset, each discharge text is structured into multiple sections. Table 8 presents the section headers keys used in this study. With the exception of the first key, "Initial", which corresponds to the opening portion of the discharge text, the remaining sections appear sequentially throughout the document. If a section key is not present in a given discharge text, we exclude that section from processing. These section keys are used to define truncation boundaries in our experiments.

Among these sections are the two target segments, BHC and DI. As specified by the task organizers, the complete discharge text was to remain intact in the released dataset. Therefore, for model development, we removed the target sections (BHC and DI) from the original text and used the remaining content as input for training and inference.

Model	Task	Avg. Flesch reading ease	Avg. sentence count	Avg. character count
Full prompt				
Gemma2	BHC	53.034	21.92	1388.392
	DI	63.876	16.932	1043.1
Llamma3	BHC	53.71	27.312	1762.892
	DI	62.596	20.888	1403.124
Mistral	BHC	49.496	18.872	1207.036
	DI	62.32	16.98	1131.912
Phi4	BHC	54.676	24.928	1520.58
	DI	60.76	18.708	1247.096
Truncated Prompt				
Gemma2	BHC	51.2	23.372	1496.2
	DI	64.028	20.524	1291.492
Llama3	BHC	53.016	26.484	1701.888
	DI	61.902	24.56	1634.052
Mistral3	BHC	52.66	21.804	1357.78
	DI	61.861	21.58	1437.784
Phi4	BHC	54.445	27.232	1672.78
	DI	61.424	22.236	1458.916

Table 4: The mean Flesch reading ease score and mean sentence and character count of finetuned model generated output

3. Readability Score

We compute readability using the Flesch Reading Ease (FRE) score [6] , which provides a numerical estimate of how easy a text is to read [4]. The FRE score ranges from 0 to 100, where higher scores indicate greater readability and lower scores indicate more complex or difficult text. The score is calculated based on average sentence length and average syllables per word.

Table 4 reports the readability scores of summaries generated by the fine-tuned model using both full and truncated input prompts. Tables 5 and 6 present the corresponding readability scores for the instruction-tuned model under a zero-shot setting, again using full and truncated inputs, respectively. Readability scores are reported separately for the BHC and DI tasks. For the zero-shot setting, results are provided for four distinct instructions.

When averaging the readability scores across tasks (BHC and DI), the fine-tuned model achieves an average FRE score of 57.5 for both full and truncated prompts. In contrast, the zero-shot model with full input yields average readability scores of 32.2, 37.7, 33.6, and 36.4 for the four instructions, while with truncated input the corresponding scores are 39.3, 39.1, 38.0, and 39.2. These results indicate that summaries generated by the fine-tuned model are consistently more readable than those produced in the zero-shot setting.

In addition to readability, we also compute the average sentence count and character count for the generated summaries. For the fine-tuned model, the average sentence counts for full and truncated inputs are 20.8 and 23.5, with corresponding character counts of 1338 and 1506.3, respectively. For the zero-shot model with full input, the average sentence counts across the four instructions are 27.7, 32.2, 31.4, and 27.0, and the character counts are 2187.0, 2458.6, 2383.8, and 2128.1, respectively. With truncated input,

[6] https://github.com/textstat/textstat?tab=readme-ov-file

Model	Task	Avg. Flesch reading ease	Avg. sentence count	Avg. character count
Gemma2 Instruction-1	BHC	14.348	7.252	653.568
	DI	39.479	23.632	2008.932
Gemma2 Instruction-2	BHC	14.779	17.396	1549.604
	DI	48.786	22.6	1707.728
Gemma2 Instruction-3	BHC	6.984	13.832	1322.628
	DI	54.597	29.828	2042.664
Gemma2 Instruction-4	BHC	10.632	9.116	914.736
	DI	47.788	18.736	1510.984
Gemma3 Instruction-1	BHC	33.247	10.352	891.596
	DI	45.602	31.58	2393.836
Gemma3 Instruction-2	BHC	31.438	21.952	1702.572
	DI	47.162	23.7	1656.624
Gemma3 Instruction-3	BHC	18.83	12.732	1269.84
	DI	52.338	31.84	2255.82
Gemma3 Instruction-4	BHC	22.578	7.724	791.192
	DI	51.433	23.156	1837.692
LLAMA3 Instruction-1	BHC	38.133	70.748	4932.8
	DI	44.571	63.704	4626.48
LLAMA3 Instruction-2	BHC	39.265	74.892	5487.888
	DI	46.154	67.588	4699.352
LLAMA3 Instruction-3	BHC	40.494	76.4	5258.3
	DI	50.308	66.204	4697.136
LLAMA3 Instruction-4	BHC	41.439	73.74	5151.684
	DI	43.853	63.792	4569.032
Mistral Instruction-1	BHC	-4.605	7.62	992.144
	DI	44.841	18.532	1538.288
Mistral Instruction-2	BHC	33.817	28.224	2291.088
	DI	47.047	19.492	1520.84
Mistral Instruction-3	BHC	21.058	13.708	1424.104
	DI	51.076	23.588	1708.776
Mistral Instruction-4	BHC	25.363	11.976	1241.64
	DI	49.597	19.244	1442.604
Phi4 Instruction-1	BHC	28.082	15.744	1607.236
	DI	38.705	27.532	2224.724
Phi4 Instruction-2	BHC	26.734	23.38	2058.32
	DI	42.189	22.628	1912.036
Phi4 Instruction-3	BHC	19.689	18.476	1806.212
	DI	47.115	27.94	2052.476
Phi4 Instruction-4	BHC	25.768	16.604	1688.868
	DI	45.936	26.244	2132.752

Table 5: The mean Flesch Reading Ease score and the average sentence count of original model-generated outputs using zero-shot full prompts with corresponding instructions.

Model	Task	Avg. Flesch reading ease	Avg. sentence count	Avg. character count
Gemma2 Instruction-1	BHC	13.851	7.332	648.93
	DI	40.194	24.298	2030.86
Gemma2 Instruction-2	BHC	14.988	17.386	1538.478
	DI	49.02	22.461	1699.801
Gemma2 Instruction-3	BHC	7.247	13.311	1268.995
	DI	54.507	30.344	2082.498
Gemma2 Instruction-4	BHC	10.851	8.948	888.122
	DI	48.504	19.085	1512.67
Gemma3 Instruction-1	BHC	35.224	11.076	929.924
	DI	48.575	23.172	1811.848
Gemma3 Instruction-2	BHC	34.287	23.412	1729.04
	DI	46.978	20.216	1456.284
Gemma3 Instruction-3	BHC	21.51	13.888	1295.556
	DI	54.583	35.58	2472.516
Gemma3 Instruction-4	BHC	25.166	9.108	874.844
	DI	52.993	19.86	1564.156
LLAMA3 Instruction-1	BHC	35.89	42.076	2826.392
	DI	43.015	26.46	2259.9
LLAMA3 Instruction-2	BHC	38.486	40.648	3322.744
	DI	47.561	23.904	1783.188
LLAMA3 Instruction-3	BHC	38.568	40.448	3038.204
	DI	53.536	28.06	2074.172
LLAMA3 Instruction-4	BHC	42.865	44.192	3261.976
	DI	44.747	19.644	1764.836
Mistral Instruction-1	BHC	41.199	17.668	1510.288
	DI	52.792	18.44	1162.732
Mistral Instruction-2	BHC	37.028	25.856	2040.572
	DI	47.704	17.284	1279.744
Mistral Instruction-3	BHC	25.469	11.74	1228.232
	DI	55.246	23.384	1565.836
Mistral Instruction-4	BHC	36.889	13.384	1292.044
	DI	52.418	20.38	1345.604
Phi4 Instruction-1	BHC	28.943	15.564	1542.56
	DI	43.704	26.756	2076.824
Phi4 Instruction-2	BHC	30.542	21.984	1911.676
	DI	44.885	22.208	1760.528
Phi4 Instruction-3	BHC	20.059	18.22	1744.372
	DI	49.714	26.572	1956.244
Phi4 Instruction-4	BHC	29.717	16.688	1649.024
	DI	48.029	25.14	1995.18

Table 6: The mean Flesch Reading Ease score and the average sentence and character count of original model-generated outputs using zero-shot Truncated prompts.

the zero-shot model produces 21.3, 23.5, 24.1 and 19.6 sentences on average, with character counts of 1680.0, 1852.2, 1872.7 and 1614.8.

Overall, the generation lengths are relatively comparable across settings; however, the longest summaries are produced by the zero-shot model with full input, while the fine-tuned model generates more concise and readable outputs.

#	Team Name	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
1	WisPerMed	**0.124**	**0.453**	**0.201**	**0.308**	**0.438**	**0.403**	**0.315**	**0.411**	**0.332**
2	HarmonAI Lab at Yale	0.106	0.423	0.18	0.284	0.412	0.381	0.265	0.353	0.3
3	aehrc	0.097	0.414	0.192	0.284	0.383	0.398	0.274	0.332	0.297
4	EPFL-MAKE	0.098	0.444	0.155	0.262	0.399	0.336	0.255	0.36	0.289
5	UF-HOBI	0.102	0.401	0.174	0.275	0.395	0.289	0.296	0.355	0.286

Table 7: Evaluation results on the Phase 2 test set show that our instruction-tuned models perform competitively with leaderboard leaders WisPerMed [5], HarmonAI Lab at Yale [6], aehrc [7], EPFL-MAKE [8], UF-HOBI [9].

#	Section Key
1	Initial
2	Allergies
3	Attending
4	Chief Complaint
5	Major Surgical or Invasive Procedure
6	History of Present Illness
7	Past Medical History
8	Social History
9	Physical Exam
10	Pertinent Results
11	Brief Hospital Course
12	Medications on Admission
13	Discharge Medications
14	Discharge Disposition
15	Discharge Diagnosis
16	Discharge Condition
17	Discharge Instructions
18	Followup Instructions

Table 8: Different sections extracted from the discharge text for the proposed task. The keys denote the corresponding sections. The key "Initial" refers to the opening portion of the discharge text, which contains the initial discharge-related information.

4. Leaderboard score

Top 5 leaderboard scores reported in the Discharge Me! shared task at the Association for Computational Linguistics BioNLP Workshop 2024 are recorded in Table 7. We only record top 5 scores and bold number are the best score corresponding to that metric.

Table 9: Different instructions for zero-shot prompting to generate "BHC" and "DI"

Zero-shot prompt Instructions for BHC and DI
BHC: Instruction 1
You are a clinical language model. Below is the summarized discharge text and radiology note. Generate the Brief Hospital Course section, focusing only on: - Clinical events - Interventions and procedures - Patient progress during admission Exclude discharge instructions and follow-up care.
DI: Instruction 1

Continuation of Table 9
You are a clinical language model. Below is the summarized discharge text and radiology note. Generate patient-facing Discharge instructions based on both texts. Include reason for admission, clinical events, interventions, discharge condition, and follow-up care.
BHC: Instruction 2
You are a clinical documentation assistant. Your task is to extract the Brief Hospital Course from a hospital discharge text and radiology notes. The Brief Hospital Course should provide a concise and structured summary of the patient's inpatient journey, including relevant history, major diagnoses, key interventions, clinical assessments, procedures performed, significant events during admission, and discharge status. Please follow the format below for a structured output: 1. Chief Complaint / Presentation: [Brief statement of the reason for admission or presenting symptoms] 2. Relevant History: [Comorbidities or relevant medical/surgical history impacting care] 3. Hospital Course Summary: 3.1. Day-by-day (or stage-by-stage) narrative of key clinical events, interventions, treatments, response to treatment, procedures (e.g., surgeries, ERCPs), and consultations. 3.2. Summarize any imaging, labs, or radiology findings only if they impacted the clinical decision-making. 3.3. Mention symptom resolution, response to therapy, and transition planning (e.g., mobility, PT, diet, discharge medications). 4. Discharge Status: [Patient condition at discharge and where they were discharged to, if available]
DI: Instruction 2
You are a clinical documentation assistant. Your task is to extract and generate the Discharge Instructions section from a patient's discharge text and radiology notes. The Discharge Instructions should be: Written in simple, clear, and patient-facing language Summarize key hospital events and explain what the patient should do after discharge Include medication changes, follow-up appointments, self-care advice, and warning signs if applicable Preserve clinical accuracy while ensuring understandability Please follow the format below for a structured output: 1. Why you came to the hospital: [Summarize the reason for admission in one sentence] 2. What happened in the hospital: [Brief summary of key findings, diagnoses, procedures, and treatment] 3. Medications: 3.1. Continue: [List important medications to continue] 3.2. Start: [Newly prescribed medications] 3.3. Stop or change: [Medications discontinued or dosage changes] 4. What to do after discharge: 4.1. [Instructions on medications, diet, mobility, wound care, etc.] 4.2. [Expected symptoms and when to seek help] 4.3. [Additional care instructions, e.g., PICC line, oxygen use] 5. Follow-up appointments: 5.1. [List scheduled follow-ups or instructions to make appointments, with providers and timeframes] 6. Other important information: 6.1 [E.g., contact numbers, documentation sent, special precautions]
BHC: Instruction 3
Generate the Brief Hospital Course section from the following discharge text, using the following structure and clinical style: 1. Begin by stating the patient,s reason for admission using clinical language. 2. Summarize the major active medical problems addressed during the hospital stay. 3. For each problem: - Describe the diagnostic workup performed (labs, imaging, consults). - Outline the treatments or interventions (medications, procedures, surgeries). - Comment on the patient,s clinical response and recovery status. 4. Include any relevant chronic issues managed or monitored during the stay. 5. Conclude with the discharge disposition (e.g., home, rehab) and any brief follow-up plans. Use concise, problem-oriented medical language appropriate for clinicians. Group content by condition using headings such as:

Continuation of Table 9
- [Medical Problem] - Chronic Conditions - Transition of Care Do not include patient instructions or non-clinical explanations. Focus only on in-hospital course and medical decision-making.
DI: Instruction 3
Generate the Discharge Instructions section from the following Discharge text, using the following structure: 1. Greet the patient and state the reason for admission. 2. Briefly explain what was done during the hospital stay, including key procedures or findings. 3. Provide detailed follow-up instructions, including: - Medication changes (new, changed, or stopped) - Lifestyle advice (diet, physical activity, restrictions) - Symptom monitoring (what to watch for, when to seek help) - Follow-up appointments and contact information 4. Use plain language that the patient can understand. 5. Maintain a compassionate, supportive tone. 6. (Optional) Use headings like: - Why did you come to the hospital? - What happened here? - What should you do now? 7. End with a warm closing message from the care team. The generated instructions should be written in a clear and empathetic style appropriate for a patient leaving the hospital.
BHC: Instruction 4
You are a clinical documentation specialist. Based on the discharge summary below, generate a concise and accurate “Brief Hospital Course” section. The section should include: - Reason for admission and presenting symptoms - Key diagnostic workup and findings - Summary of treatment provided during admission - Patient’s response to treatment - Discharge condition and disposition Write clearly in clinical style suitable for inclusion in a discharge summary. Use paragraph format or structured bullets.
DI: Instruction 4
You are a clinical documentation assistant. Read the full discharge summary provided below and generate a clear, patient-facing “Discharge Instructions” section. Instructions should: - Begin with a statement of why the patient was admitted. - Mention whether the cause was identified, and summarize relevant findings or evaluations. - Include any major tests, procedures, or consultations that occurred or are planned. - Recommend follow-up care and outpatient appointments. - List the medications prescribed at discharge for symptom management. - Use plain language suitable for patients or caregivers. - Keep the tone supportive and informative.

5. Instruction sets for BHC and DI

For our task, we use four different sets of instructions, as shown in Table 9. These prompts were originally designed for zero-shot prompting. For each instruction set, we include separate prompts for generating the BHC and DI sections. The first instruction set is additionally used for fine-tuning the models.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.093	0.402	0.173	0.281	0.397	0.273	0.324	0.425	0.296
700	0.113	0.414	0.178	**0.284**	0.4	0.31	0.308	0.44	0.306
900	**0.116**	0.411	**0.179**	0.281	0.397	0.329	0.297	**0.445***	**0.307**
1100	0.111	0.402	0.175	0.275	0.391	0.337	0.293	0.441	0.303
1300	0.107	0.393	0.172	0.269	0.385	0.339	0.289	0.436	0.299
1500	0.103	0.385	0.169	0.264	0.381	0.34	0.287	0.431	0.295
1700	0.101	0.379	0.167	0.26	0.378	0.339	0.285	0.43	0.292
1900	0.1	0.374	0.165	0.257	0.376	0.339	0.283	0.426	0.29
2100	0.099	0.371	0.164	0.255	0.374	0.338	0.282	0.423	0.288
Llama3									
500	0.086	0.404	0.167	0.273	0.396	0.272	0.309	0.417	0.29
700	0.108	**0.424**	0.176	0.279	**0.402**	0.319	0.296	0.437	0.305
900	0.109	0.421	0.176	0.274	0.398	0.342	0.286	0.443	0.306
1100	0.101	0.408	0.169	0.265	0.387	0.35	0.28	0.436	0.3
1300	0.092	0.393	0.162	0.255	0.377	**0.352**	0.274	0.426	0.291
1500	0.086	0.381	0.157	0.246	0.369	**0.352**	0.271	0.417	0.285
1700	0.081	0.372	0.152	0.24	0.364	0.35	0.268	0.41	0.28
1900	0.078	0.362	0.148	0.234	0.359	0.348	0.265	0.404	0.275
2100	0.075	0.355	0.145	0.229	0.356	0.346	0.262	0.398	0.271
Mistral									
500	0.092	0.392	0.169	0.273	0.389	0.267	0.321	0.42	0.29
700	0.112	0.404	0.176	0.277	0.392	0.307	0.306	0.433	0.301
900	0.112	0.399	0.174	0.271	0.387	0.324	0.298	0.436	0.3
1100	0.107	0.389	0.17	0.265	0.381	0.331	0.293	0.43	0.296
1300	0.101	0.378	0.165	0.257	0.373	0.33	0.288	0.424	0.289
1500	0.096	0.37	0.161	0.252	0.368	0.331	0.284	0.419	0.285
1700	0.093	0.364	0.159	0.247	0.365	0.33	0.282	0.415	0.282
1900	0.092	0.36	0.157	0.244	0.363	0.329	0.28	0.412	0.28
2100	0.09	0.356	0.156	0.242	0.362	0.329	0.278	0.408	0.278
Phi4									
500	0.072	0.404	0.173	0.279	0.375	0.275	**0.396***	0.429	0.3
700	0.086	0.413	0.177	0.28	0.373	0.309	0.38	0.436	**0.307**
900	0.087	0.406	0.174	0.274	0.366	0.323	0.367	0.428	0.303
1100	0.082	0.396	0.17	0.266	0.359	0.329	0.357	0.421	0.298
1300	0.078	0.386	0.167	0.259	0.353	0.331	0.352	0.415	0.293
1500	0.074	0.378	0.163	0.253	0.348	0.332	0.347	0.41	0.288
1700	0.072	0.371	0.161	0.249	0.346	0.331	0.344	0.405	0.285
1900	0.072	0.366	0.16	0.247	0.345	0.331	0.342	0.401	0.283
2100	0.07	0.361	0.158	0.244	0.343	0.33	0.34	0.395	0.28

Table 10: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the full prompt. Bold values indicate the highest score in each column. We additionally compare our results with the official performance scores of other teams reported in the Discharge Me! shared task at the Association for Computational Linguistics BioNLP Workshop 2024. An asterisk (*) denotes cases where our score surpasses the corresponding scores of all other participating teams.

6. Variation of scores over generation length

Table 10 reports the evaluation scores computed over the first *n* characters of summaries generated by the fine-tuned models. Tables 13–19 present the corresponding scores for zero-shot prompting with truncated input, using the four different instruction

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.088	0.391	0.162	0.269	0.384	0.265	0.3	0.398	0.282
700	0.103	0.399	0.164	0.27	0.384	0.297	0.288	0.411	0.29
900	0.102	0.393	0.163	0.264	0.377	0.314	0.281	0.412	0.288
1100	0.096	0.382	0.158	0.256	0.37	0.32	0.273	0.41	0.283
1300	0.09	0.371	0.154	0.248	0.363	0.322	0.269	0.408	0.278
1500	0.087	0.362	0.151	0.242	0.358	0.323	0.266	0.402	0.274
1700	0.084	0.353	0.148	0.237	0.354	0.323	0.263	0.398	0.27
1900	0.083	0.347	0.146	0.233	0.352	0.322	0.261	0.394	0.267
2100	0.081	0.342	0.144	0.23	0.35	0.321	0.258	0.389	0.264
Llama3									
500	0.079	0.391	0.154	0.261	0.382	0.261	0.296	0.411	0.279
700	0.097	**0.407**	0.161	0.266	0.385	0.304	0.283	**0.434***	0.292
900	0.098	0.403	0.159	0.259	0.378	0.325	0.273	0.432	0.291
1100	0.09	0.388	0.152	0.248	0.367	0.332	0.264	0.423	0.283
1300	0.083	0.374	0.146	0.239	0.357	**0.335**	0.261	0.413	0.276
1500	0.077	0.362	0.141	0.231	0.349	0.335	0.257	0.408	0.27
1700	0.074	0.352	0.138	0.225	0.344	0.334	0.254	0.403	0.266
1900	0.071	0.344	0.135	0.221	0.341	0.334	0.252	0.397	0.262
2100	0.069	0.338	0.133	0.217	0.338	0.333	0.25	0.391	0.259
Mistral									
500	0.094	0.393	0.169	0.273	**0.388**	0.268	**0.313**	0.407	0.288
700	**0.111**	0.402	**0.172**	**0.273**	0.387	0.305	0.298	0.419	**0.296**
900	0.11	0.396	0.17	0.267	0.381	0.325	0.285	0.419	0.294
1100	0.103	0.383	0.164	0.258	0.372	0.33	0.276	0.412	0.287
1300	0.096	0.371	0.158	0.249	0.364	0.331	0.271	0.406	0.281
1500	0.091	0.36	0.153	0.241	0.357	0.33	0.267	0.401	0.275
1700	0.087	0.351	0.149	0.235	0.352	0.328	0.264	0.395	0.27
1900	0.084	0.344	0.146	0.231	0.349	0.326	0.262	0.391	0.267
2100	0.082	0.338	0.144	0.227	0.347	0.324	0.26	0.386	0.263
Phi4									
500	0.087	0.388	0.162	0.269	0.382	0.262	0.298	0.393	0.28
700	0.104	0.398	0.165	0.271	0.383	0.3	0.283	0.406	0.289
900	0.105	0.39	0.162	0.264	0.377	0.315	0.275	0.406	0.287
1100	0.099	0.379	0.158	0.256	0.37	0.324	0.268	0.401	0.282
1300	0.094	0.366	0.154	0.248	0.362	0.326	0.266	0.396	0.277
1500	0.09	0.356	0.15	0.241	0.356	0.325	0.263	0.392	0.272
1700	0.087	0.347	0.147	0.235	0.351	0.325	0.26	0.389	0.268
1900	0.086	0.34	0.145	0.231	0.348	0.324	0.259	0.383	0.264
2100	0.084	0.335	0.143	0.227	0.346	0.323	0.255	0.379	0.262

Table 11: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the truncated prompt. Bold values indicate the highest score in each column. We additionally compare our results with the official performance scores of other teams reported in the Discharge Me! shared task at the Association for Computational Linguistics BioNLP Workshop 2024. An asterisk (*) denotes cases where our score surpasses the corresponding scores of all other participating teams.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.012	0.265	0.057	0.151	0.239	0.159	0.24	0.302	0.178
700	0.013	0.285	0.059	0.158	0.251	0.188	0.237	0.314	0.188
900	0.012	0.29	0.058	0.159	0.253	0.205	0.237	0.309	0.19
1100	0.011	0.289	0.058	0.157	0.252	0.213	0.236	0.301	0.19
1300	0.011	0.285	0.057	0.155	0.25	0.218	0.234	0.297	0.188
1500	0.01	0.28	0.056	0.152	0.247	0.219	0.232	0.294	0.186
1700	0.01	0.275	0.055	0.149	0.245	0.22	0.231	0.289	0.184
1900	0.009	0.27	0.054	0.147	0.243	0.22	0.23	0.285	0.182
2100	0.009	0.267	0.054	0.144	0.242	0.219	0.23	0.281	0.181
Gemma3									
500	0.02	0.296	0.072	0.167	0.247	0.178	0.243	0.302	0.191
700	**0.023**	0.315	0.074	0.174	0.26	0.214	0.229	**0.318**	0.201
900	0.022	**0.32**	**0.075**	**0.175**	**0.264**	0.237	0.223	**0.318**	**0.204**
1100	0.021	0.319	**0.075**	0.174	**0.264**	0.251	0.217	0.311	**0.204**
1300	0.02	0.314	0.074	0.171	0.26	0.256	0.215	0.303	0.202
1500	0.019	0.308	0.073	0.168	0.255	0.258	0.215	0.299	0.199
1700	0.019	0.302	0.072	0.165	0.25	0.259	0.214	0.291	0.196
1900	0.018	0.296	0.071	0.162	0.248	0.259	0.215	0.287	0.195
2100	0.018	0.292	0.07	0.159	0.247	0.259	0.216	0.284	0.193
LLAMA3									
500	0.007	0.232	0.039	0.132	0.185	0.142	**0.284**	0.235	0.157
700	0.009	0.244	0.042	0.133	0.187	0.166	0.267	0.26	0.163
900	0.01	0.253	0.043	0.134	0.194	0.188	0.267	0.273	0.17
1100	0.01	0.253	0.044	0.133	0.197	0.203	0.26	0.27	0.171
1300	0.009	0.248	0.044	0.129	0.194	0.213	0.254	0.268	0.17
1500	0.009	0.241	0.043	0.126	0.191	0.218	0.253	0.26	0.168
1700	0.009	0.234	0.042	0.122	0.189	0.222	0.252	0.256	0.166
1900	0.008	0.226	0.041	0.119	0.186	0.223	0.253	0.252	0.163
2100	0.008	0.219	0.041	0.116	0.183	0.224	0.253	0.248	0.161
Mistral									
500	0.012	0.268	0.058	0.158	0.232	0.16	0.208	0.255	0.169
700	0.014	0.284	0.06	0.161	0.233	0.191	0.212	0.262	0.177
900	0.014	0.288	0.06	0.16	0.231	0.209	0.212	0.263	0.18
1100	0.014	0.287	0.061	0.158	0.229	0.222	0.215	0.263	0.181
1300	0.014	0.284	0.061	0.155	0.227	0.231	0.22	0.263	0.182
1500	0.014	0.281	0.061	0.153	0.226	0.235	0.222	0.261	0.182
1700	0.014	0.279	0.061	0.151	0.225	0.238	0.221	0.259	0.181
1900	0.014	0.278	0.061	0.15	0.225	0.24	0.223	0.259	0.181
2100	0.014	0.277	0.061	0.15	0.225	0.24	0.222	0.258	0.181
Phi4									
500	0.017	0.281	0.063	0.166	0.239	0.167	0.226	0.28	0.18
700	0.019	0.301	0.065	0.169	0.248	0.201	0.214	0.293	0.189
900	0.019	0.304	0.065	0.168	0.25	0.221	0.206	0.293	0.191
1100	0.017	0.303	0.065	0.165	0.247	0.237	0.204	0.289	0.191
1300	0.016	0.296	0.063	0.159	0.24	0.247	0.207	0.282	0.189
1500	0.015	0.29	0.063	0.155	0.236	0.255	0.212	0.278	0.188
1700	0.014	0.284	0.063	0.153	0.234	0.261	0.216	0.278	0.188
1900	0.014	0.279	0.063	0.15	0.234	0.265	0.217	0.275	0.187
2100	0.014	0.274	0.063	0.148	0.234	**0.266**	0.22	0.273	0.187

Table 12: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the full prompt 1. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.017	0.301	**0.073**	**0.181**	**0.256**	0.171	**0.289**	0.301	**0.199**
700	0.017	0.299	0.071	0.175	0.243	0.186	0.273	0.293	0.195
900	0.016	0.293	0.069	0.17	0.234	0.193	0.266	0.287	0.191
1100	0.014	0.287	0.068	0.166	0.229	0.196	0.261	0.283	0.188
1300	0.013	0.284	0.067	0.162	0.227	0.2	0.259	0.28	0.186
1500	0.013	0.281	0.067	0.16	0.227	0.202	0.257	0.28	0.186
1700	0.013	0.278	0.066	0.158	0.227	0.202	0.257	0.279	0.185
1900	0.013	0.277	0.066	0.157	0.227	0.204	0.258	0.279	0.185
2100	0.013	0.276	0.066	0.156	0.227	0.204	0.258	0.279	0.185
Gemma3									
500	0.016	0.3	0.068	0.164	0.238	0.181	0.274	0.292	0.192
700	0.018	**0.312**	0.069	0.169	0.246	0.212	0.254	**0.309**	**0.199**
900	0.018	**0.312**	0.068	0.167	0.244	0.229	0.246	0.303	0.198
1100	0.017	0.305	0.066	0.162	0.236	0.236	0.239	0.295	0.195
1300	0.016	0.296	0.064	0.158	0.229	0.239	0.236	0.287	0.191
1500	0.016	0.29	0.064	0.155	0.226	0.24	0.236	0.283	0.189
1700	0.015	0.286	0.064	0.153	0.224	0.242	0.238	0.281	0.188
1900	0.015	0.282	0.063	0.152	0.223	0.241	0.239	0.279	0.187
2100	0.015	0.279	0.063	0.15	0.223	0.24	0.241	0.279	0.186
LLAMA3									
500	0.012	0.278	0.055	0.154	0.228	0.163	0.283	0.268	0.18
700	0.013	0.295	0.057	0.157	0.233	0.192	0.275	0.282	0.188
900	0.013	0.294	0.057	0.157	0.23	0.21	0.272	0.279	0.189
1100	0.013	0.288	0.056	0.153	0.225	0.222	0.27	0.274	0.188
1300	0.012	0.278	0.055	0.149	0.218	0.228	0.272	0.273	0.186
1500	0.011	0.268	0.054	0.144	0.213	0.233	0.273	0.27	0.183
1700	0.011	0.258	0.053	0.14	0.21	0.234	0.274	0.267	0.181
1900	0.01	0.249	0.051	0.136	0.207	0.234	0.276	0.265	0.178
2100	0.01	0.242	0.05	0.132	0.203	0.233	0.277	0.263	0.176
Mistral									
500	0.016	0.264	0.06	0.16	0.213	0.167	0.2	0.245	0.166
700	0.018	0.276	0.061	0.159	0.208	0.195	0.196	0.251	0.171
900	**0.019**	0.281	0.06	0.156	0.21	0.216	0.196	0.256	0.174
1100	0.018	0.281	0.061	0.153	0.211	0.231	0.198	0.257	0.176
1300	0.017	0.28	0.06	0.152	0.212	0.241	0.199	0.254	0.177
1500	0.017	0.278	0.061	0.15	0.213	0.247	0.201	0.252	0.177
1700	0.017	0.276	0.061	0.149	0.214	0.251	0.202	0.25	0.177
1900	0.016	0.275	0.061	0.149	0.214	0.252	0.202	0.248	0.177
2100	0.016	0.275	0.061	0.149	0.214	**0.253**	0.202	0.247	0.177
Phi4									
500	0.015	0.281	0.06	0.165	0.244	0.166	0.285	0.273	0.186
700	0.017	0.297	0.061	0.168	0.248	0.197	0.259	0.284	0.192
900	0.018	0.304	0.062	0.168	0.248	0.221	0.243	0.288	0.194
1100	0.016	0.299	0.061	0.161	0.239	0.233	0.234	0.282	0.191
1300	0.015	0.289	0.059	0.155	0.228	0.24	0.234	0.276	0.187
1500	0.014	0.28	0.058	0.15	0.222	0.244	0.238	0.271	0.185
1700	0.013	0.273	0.057	0.146	0.219	0.247	0.24	0.27	0.183
1900	0.012	0.267	0.057	0.143	0.219	0.249	0.243	0.268	0.182
2100	0.012	0.264	0.057	0.142	0.218	0.25	0.245	0.267	0.182

Table 13: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the truncated prompt 1. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.006	0.243	0.05	0.139	0.219	0.16	0.275	0.285	0.172
700	0.008	0.261	0.051	0.144	0.22	0.186	0.265	**0.291**	0.178
900	0.008	0.271	0.052	0.145	0.219	0.207	0.259	0.29	0.181
1100	0.008	0.275	0.051	0.145	0.217	0.221	0.255	0.288	0.183
1300	0.008	0.275	0.051	0.142	0.216	0.231	0.256	0.284	0.183
1500	0.008	0.275	0.051	0.141	0.216	0.239	0.256	0.282	0.183
1700	0.008	0.274	0.052	0.14	0.217	0.245	0.259	0.278	0.184
1900	0.008	0.272	0.052	0.138	0.218	0.248	0.26	0.275	0.184
2100	0.008	0.271	0.053	0.138	0.219	0.25	0.26	0.274	0.184
Gemma3									
500	0.015	0.268	**0.066**	0.155	**0.238**	0.175	0.261	0.29	0.184
700	0.016	0.281	0.065	**0.157**	0.235	0.202	0.248	0.288	0.187
900	**0.017**	0.284	0.065	0.154	0.227	0.22	0.246	**0.291**	0.188
1100	**0.017**	**0.285**	0.065	0.151	0.223	0.236	0.245	0.289	**0.189**
1300	0.016	0.284	0.064	0.15	0.218	0.247	0.247	0.284	**0.189**
1500	0.016	0.281	0.064	0.147	0.218	0.256	0.249	0.279	**0.189**
1700	0.015	0.277	0.064	0.145	0.218	0.262	0.251	0.274	0.188
1900	0.015	0.275	0.064	0.144	0.22	0.266	0.252	0.271	0.188
2100	0.014	0.272	0.064	0.143	0.22	**0.268**	0.252	0.27	0.188
LLAMA3									
500	0.01	0.235	0.046	0.139	0.193	0.15	0.274	0.24	0.161
700	0.012	0.247	0.047	0.14	0.189	0.175	0.277	0.248	0.167
900	0.013	0.252	0.048	0.138	0.183	0.195	0.281	0.257	0.171
1100	0.013	0.253	0.048	0.136	0.177	0.207	0.275	0.263	0.171
1300	0.012	0.251	0.048	0.134	0.174	0.217	0.285	0.266	0.173
1500	0.011	0.247	0.047	0.131	0.173	0.225	0.283	0.266	0.173
1700	0.011	0.242	0.047	0.127	0.172	0.229	0.285	0.267	0.173
1900	0.01	0.238	0.047	0.124	0.175	0.234	**0.286**	0.266	0.173
2100	0.01	0.233	0.046	0.121	0.176	0.237	0.285	0.263	0.171
Mistral									
500	0.011	0.233	0.05	0.137	0.193	0.152	0.245	0.225	0.156
700	0.013	0.249	0.051	0.139	0.195	0.18	0.239	0.236	0.163
900	0.014	0.259	0.053	0.139	0.193	0.202	0.235	0.249	0.168
1100	0.015	0.263	0.054	0.138	0.189	0.218	0.237	0.256	0.171
1300	0.014	0.263	0.055	0.136	0.187	0.23	0.236	0.258	0.172
1500	0.013	0.261	0.054	0.134	0.187	0.239	0.239	0.258	0.173
1700	0.013	0.257	0.054	0.133	0.187	0.244	0.239	0.257	0.173
1900	0.012	0.254	0.054	0.131	0.187	0.247	0.24	0.254	0.172
2100	0.012	0.25	0.054	0.129	0.187	0.249	0.241	0.25	0.171
Phi4									
500	0.008	0.258	0.056	0.143	0.213	0.165	0.251	0.265	0.17
700	0.01	0.273	0.059	0.148	0.216	0.194	0.249	0.266	0.177
900	0.011	0.277	0.059	0.146	0.208	0.211	0.243	0.267	0.178
1100	0.012	0.277	0.059	0.144	0.202	0.225	0.238	0.269	0.178
1300	0.012	0.278	0.059	0.142	0.201	0.236	0.236	0.268	0.179
1500	0.011	0.277	0.06	0.141	0.201	0.245	0.237	0.267	0.18
1700	0.011	0.275	0.06	0.14	0.202	0.253	0.237	0.265	0.18
1900	0.011	0.271	0.06	0.139	0.204	0.257	0.239	0.261	0.18
2100	0.011	0.267	0.06	0.138	0.205	0.261	0.243	0.26	0.181

Table 14: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the full prompt 2. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.011	0.254	0.059	0.148	0.207	0.168	0.265	0.261	0.172
700	0.014	0.277	0.062	0.153	0.209	0.2	0.263	0.274	0.182
900	0.016	0.291	0.065	0.157	0.214	0.223	0.265	0.287	0.19
1100	0.016	0.296	0.066	**0.158**	0.217	0.236	0.268	**0.291**	0.194
1300	0.016	**0.298**	**0.067**	**0.158**	0.218	0.241	0.271	**0.291**	0.195
1500	0.016	**0.298**	**0.067**	**0.158**	0.219	0.245	0.271	0.29	**0.196**
1700	0.016	**0.298**	**0.067**	**0.158**	**0.22**	0.246	0.272	0.29	**0.196**
1900	0.016	**0.298**	**0.067**	**0.158**	**0.22**	0.246	0.272	0.29	**0.196**
2100	0.016	**0.298**	**0.067**	**0.158**	**0.22**	0.246	0.272	0.29	**0.196**
Gemma3									
500	0.012	0.257	0.06	0.148	0.212	0.166	**0.291**	0.257	0.176
700	0.015	0.268	0.06	0.148	0.205	0.19	0.281	0.264	0.179
900	0.016	0.274	0.061	0.148	0.203	0.211	0.274	0.271	0.182
1100	**0.017**	0.277	0.061	0.147	0.201	0.225	0.269	0.274	0.184
1300	0.016	0.276	0.061	0.144	0.199	0.236	0.266	0.272	0.184
1500	0.015	0.275	0.06	0.143	0.199	0.243	0.267	0.268	0.184
1700	0.014	0.271	0.059	0.141	0.2	0.246	0.267	0.265	0.183
1900	0.014	0.269	0.059	0.139	0.199	0.249	0.269	0.264	0.183
2100	0.014	0.267	0.059	0.138	0.199	0.251	0.27	0.262	0.182
LLAMA3									
500	0.011	0.251	0.056	0.147	0.207	0.164	0.27	0.253	0.17
700	0.013	0.261	0.055	0.145	0.198	0.188	0.268	0.259	0.173
900	0.014	0.27	0.057	0.146	0.198	0.21	0.274	0.264	0.179
1100	0.015	0.274	0.058	0.146	0.198	0.225	0.274	0.271	0.183
1300	0.014	0.274	0.059	0.145	0.199	0.238	0.276	0.273	0.185
1500	0.013	0.27	0.058	0.142	0.198	0.244	0.276	0.273	0.184
1700	0.012	0.266	0.057	0.14	0.199	0.248	0.277	0.27	0.184
1900	0.012	0.261	0.057	0.138	0.199	0.251	0.278	0.269	0.183
2100	0.011	0.256	0.056	0.137	0.198	**0.252**	0.278	0.268	0.182
Mistral									
500	0.012	0.237	0.056	0.142	0.193	0.158	0.256	0.236	0.161
700	0.014	0.248	0.055	0.141	0.184	0.181	0.255	0.245	0.165
900	0.015	0.261	0.056	0.142	0.184	0.204	0.251	0.255	0.171
1100	0.015	0.266	0.058	0.141	0.185	0.222	0.248	0.265	0.175
1300	0.014	0.267	0.059	0.14	0.186	0.233	0.246	0.267	0.177
1500	0.014	0.264	0.058	0.138	0.186	0.239	0.246	0.268	0.177
1700	0.013	0.263	0.058	0.137	0.187	0.245	0.247	0.266	0.177
1900	0.013	0.26	0.057	0.136	0.186	0.247	0.248	0.264	0.176
2100	0.013	0.257	0.057	0.134	0.186	0.248	0.248	0.261	0.176
Phi4									
500	0.01	0.251	0.056	0.144	0.203	0.162	0.282	0.246	0.169
700	0.011	0.257	0.055	0.143	0.192	0.183	0.276	0.244	0.17
900	0.013	0.265	0.056	0.143	0.187	0.202	0.267	0.247	0.172
1100	0.012	0.271	0.056	0.142	0.187	0.217	0.262	0.253	0.175
1300	0.012	0.272	0.057	0.141	0.186	0.228	0.258	0.261	0.177
1500	0.012	0.272	0.057	0.139	0.186	0.236	0.257	0.263	0.178
1700	0.012	0.269	0.057	0.137	0.187	0.242	0.257	0.259	0.177
1900	0.011	0.265	0.056	0.136	0.187	0.245	0.26	0.258	0.177
2100	0.011	0.261	0.056	0.135	0.187	0.247	0.263	0.258	0.177

Table 15: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the truncated prompt 2. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.004	0.255	0.045	0.141	0.219	0.158	0.248	0.301	0.171
700	0.007	0.276	0.049	0.147	0.234	0.19	0.252	**0.318**	0.184
900	0.007	0.286	0.05	0.15	0.24	0.214	0.253	0.317	0.189
1100	0.007	0.289	0.05	0.149	0.242	0.229	0.248	0.312	0.191
1300	0.007	0.286	0.05	0.147	0.241	0.237	0.241	0.303	0.189
1500	0.007	0.281	0.05	0.143	0.238	0.243	0.235	0.296	0.187
1700	0.006	0.276	0.049	0.141	0.235	0.244	0.231	0.291	0.184
1900	0.006	0.271	0.049	0.138	0.234	0.245	0.229	0.287	0.183
2100	0.006	0.267	0.049	0.136	0.233	0.245	0.228	0.284	0.181
Gemma3									
500	0.012	0.285	0.063	0.165	0.261	0.171	0.274	0.304	0.192
700	**0.013**	0.304	**0.064**	0.17	**0.266**	0.205	0.267	0.312	0.2
900	**0.013**	**0.309**	0.063	0.169	0.265	0.226	0.256	0.313	**0.202**
1100	**0.013**	0.307	0.063	0.165	0.259	0.24	0.246	0.305	0.2
1300	0.012	0.302	0.062	0.161	0.254	0.25	0.24	0.296	0.197
1500	0.012	0.298	0.062	0.159	0.251	0.257	0.237	0.29	0.196
1700	0.012	0.293	0.062	0.156	0.25	0.26	0.235	0.289	0.195
1900	0.011	0.289	0.062	0.154	0.251	0.262	0.235	0.286	0.194
2100	0.012	0.285	0.062	0.152	0.251	0.262	0.236	0.284	0.193
LLAMA3									
500	0.007	0.236	0.039	0.131	0.18	0.147	**0.298**	0.246	0.16
700	0.008	0.252	0.042	0.136	0.191	0.178	0.293	0.271	0.171
900	0.009	0.256	0.044	0.137	0.191	0.196	0.288	0.273	0.174
1100	0.009	0.257	0.044	0.135	0.19	0.209	0.285	0.269	0.175
1300	0.008	0.254	0.044	0.132	0.187	0.218	0.277	0.271	0.174
1500	0.008	0.247	0.044	0.127	0.183	0.223	0.279	0.27	0.173
1700	0.008	0.244	0.045	0.125	0.186	0.231	0.279	0.27	0.174
1900	0.008	0.237	0.044	0.121	0.187	0.232	0.277	0.263	0.171
2100	0.007	0.229	0.043	0.117	0.185	0.233	0.275	0.258	0.168
Mistral									
500	0.009	0.259	0.052	0.15	0.232	0.159	0.249	0.251	0.17
700	0.01	0.273	0.053	0.151	0.231	0.188	0.244	0.259	0.176
900	0.01	0.276	0.053	0.149	0.225	0.208	0.234	0.258	0.177
1100	0.01	0.274	0.052	0.145	0.218	0.22	0.232	0.259	0.176
1300	0.01	0.271	0.052	0.142	0.213	0.23	0.231	0.255	0.175
1500	0.009	0.267	0.052	0.139	0.211	0.236	0.231	0.255	0.175
1700	0.009	0.265	0.052	0.137	0.211	0.241	0.234	0.254	0.175
1900	0.009	0.262	0.053	0.135	0.211	0.244	0.236	0.256	0.176
2100	0.009	0.261	0.053	0.135	0.212	0.246	0.239	0.255	0.176
Phi4									
500	0.009	0.278	0.058	0.156	0.242	0.167	0.258	0.274	0.18
700	0.011	0.296	0.06	0.161	0.252	0.2	0.26	0.295	0.192
900	0.011	0.305	0.061	0.163	0.254	0.224	0.258	0.3	0.197
1100	0.011	0.305	0.06	0.161	0.248	0.239	0.25	0.295	0.196
1300	0.01	0.298	0.059	0.156	0.239	0.248	0.241	0.285	0.192
1500	0.009	0.292	0.057	0.152	0.234	0.254	0.238	0.28	0.19
1700	0.009	0.285	0.057	0.149	0.231	0.258	0.239	0.277	0.188
1900	0.009	0.279	0.058	0.146	0.231	0.261	0.242	0.276	0.188
2100	0.009	0.274	0.058	0.143	0.232	**0.264**	0.245	0.275	0.188

Table 16: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the full prompt 3. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.007	0.285	0.058	0.163	0.24	0.17	0.294	0.309	0.191
700	0.01	0.302	0.061	**0.167**	0.243	0.201	0.281	0.309	0.197
900	0.011	**0.306**	0.062	0.166	0.241	0.221	0.269	0.299	0.197
1100	0.012	0.304	**0.063**	0.163	0.237	0.232	0.259	0.292	0.195
1300	0.012	0.301	**0.063**	0.161	0.235	0.237	0.251	0.291	0.194
1500	0.012	0.297	**0.063**	0.159	0.235	0.24	0.247	0.291	0.193
1700	0.012	0.294	**0.063**	0.157	0.235	0.24	0.246	0.291	0.192
1900	0.012	0.291	**0.063**	0.155	0.235	0.242	0.245	0.291	0.192
2100	0.012	0.29	**0.063**	0.155	0.236	0.242	0.245	0.291	0.192
Gemma3									
500	0.007	0.282	0.056	0.161	0.248	0.168	0.296	0.299	0.19
700	0.009	0.298	0.058	0.166	**0.251**	0.199	0.287	0.311	0.197
900	0.01	0.302	0.059	0.165	0.249	0.221	0.273	**0.312**	**0.199**
1100	0.01	0.298	0.058	0.161	0.242	0.232	0.262	0.302	0.195
1300	0.009	0.291	0.056	0.155	0.235	0.238	0.254	0.294	0.192
1500	0.009	0.286	0.056	0.151	0.23	0.243	0.251	0.288	0.189
1700	0.008	0.28	0.055	0.149	0.228	0.244	0.249	0.283	0.187
1900	0.008	0.276	0.055	0.146	0.226	0.244	0.249	0.282	0.186
2100	0.008	0.272	0.054	0.144	0.225	0.245	0.25	0.281	0.185
LLAMA 3									
500	0.009	0.271	0.052	0.151	0.222	0.166	0.265	0.284	0.177
700	0.011	0.287	0.055	0.157	0.232	0.197	0.265	0.293	0.187
900	0.012	0.292	0.056	0.157	0.232	0.217	0.264	0.293	0.19
1100	0.011	0.288	0.055	0.154	0.226	0.228	0.264	0.287	0.189
1300	0.011	0.282	0.054	0.15	0.22	0.236	0.261	0.285	0.188
1500	0.01	0.277	0.054	0.147	0.217	0.243	0.261	0.283	0.186
1700	0.01	0.269	0.054	0.144	0.215	0.247	0.261	0.279	0.185
1900	0.01	0.26	0.053	0.14	0.213	0.248	0.262	0.276	0.183
2100	0.01	0.252	0.052	0.136	0.21	0.248	0.263	0.274	0.181
Mistral									
500	0.011	0.262	0.052	0.151	0.228	0.162	0.271	0.265	0.175
700	0.013	0.271	0.053	0.152	0.222	0.188	0.256	0.262	0.177
900	0.013	0.274	0.053	0.149	0.215	0.205	0.249	0.259	0.177
1100	0.013	0.275	0.053	0.147	0.212	0.218	0.244	0.259	0.177
1300	0.012	0.275	0.053	0.144	0.21	0.227	0.242	0.262	0.178
1500	0.011	0.275	0.053	0.143	0.211	0.235	0.239	0.263	0.179
1700	0.011	0.273	0.054	0.142	0.213	0.24	0.24	0.263	0.18
1900	0.011	0.273	0.054	0.142	0.214	0.243	0.241	0.263	0.18
2100	0.011	0.271	0.055	0.142	0.214	0.244	0.242	0.262	0.18
Phi4									
500	0.014	0.272	0.058	0.154	0.235	0.165	0.315	0.269	0.185
700	**0.016**	0.289	0.059	0.161	0.243	0.196	**0.32**	0.284	0.196
900	0.015	0.299	0.059	0.161	0.242	0.218	0.307	0.285	0.198
1100	0.014	0.297	0.058	0.158	0.234	0.231	0.292	0.281	0.196
1300	0.013	0.293	0.057	0.154	0.226	0.239	0.281	0.273	0.192
1500	0.012	0.288	0.056	0.15	0.221	0.244	0.275	0.271	0.19
1700	0.011	0.283	0.057	0.147	0.22	0.249	0.275	0.27	0.189
1900	0.011	0.278	0.057	0.145	0.22	0.253	0.276	0.271	0.189
2100	0.011	0.272	0.058	0.143	0.22	**0.254**	0.277	0.271	0.188

Table 17: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the truncated prompt 3. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.013	0.269	0.058	0.155	0.25	0.162	0.24	0.312	0.183
700	0.015	0.292	0.06	0.163	0.264	0.197	0.232	**0.325**	0.194
900	0.015	0.301	0.061	0.165	0.268	0.22	0.229	**0.325**	0.198
1100	0.014	0.302	0.06	0.162	0.267	0.232	0.225	0.317	0.197
1300	0.013	0.299	0.06	0.16	0.264	0.239	0.221	0.312	0.196
1500	0.013	0.295	0.059	0.157	0.262	0.24	0.219	0.307	0.194
1700	0.012	0.292	0.059	0.155	0.26	0.241	0.219	0.304	0.193
1900	0.012	0.29	0.059	0.154	0.26	0.241	0.219	0.302	0.192
2100	0.012	0.288	0.058	0.153	0.259	0.241	0.218	0.301	0.191
Gemma3									
500	0.018	0.294	0.072	0.171	0.26	0.178	0.242	0.305	0.193
700	0.02	0.316	0.076	0.179	0.272	0.215	0.234	0.314	0.203
900	**0.021**	**0.326**	0.078	**0.183**	0.278	0.239	0.232	0.322	0.21
1100	**0.021**	0.325	**0.079**	0.181	0.28	0.25	0.229	0.323	**0.211**
1300	0.02	0.322	0.078	0.178	0.278	0.256	0.227	0.32	0.21
1500	0.019	0.316	0.077	0.174	0.273	0.257	0.226	0.315	0.207
1700	0.018	0.31	0.076	0.17	0.269	0.258	0.225	0.309	0.204
1900	0.018	0.305	0.074	0.167	0.265	0.257	0.226	0.304	0.202
2100	0.018	0.301	0.074	0.165	0.263	0.257	0.226	0.3	0.2
LLAMA3									
500	0.013	0.261	0.051	0.15	0.206	0.161	**0.272**	0.26	0.172
700	0.015	0.276	0.053	0.153	0.215	0.189	0.259	0.277	0.18
900	0.015	0.276	0.053	0.15	0.216	0.206	0.258	0.284	0.182
1100	0.015	0.274	0.054	0.146	0.215	0.22	0.257	0.286	0.183
1300	0.014	0.27	0.055	0.142	0.215	0.233	0.253	0.282	0.183
1500	0.014	0.263	0.054	0.138	0.213	0.24	0.249	0.272	0.18
1700	0.013	0.254	0.053	0.134	0.209	0.244	0.246	0.266	0.177
1900	0.012	0.245	0.052	0.13	0.207	0.247	0.246	0.262	0.175
2100	0.012	0.237	0.05	0.126	0.204	0.246	0.247	0.257	0.173
Mistral									
500	0.016	0.276	0.061	0.162	0.236	0.167	0.211	0.26	0.174
700	0.018	0.289	0.061	0.163	0.24	0.196	0.211	0.276	0.182
900	0.018	0.294	0.062	0.162	0.238	0.219	0.206	0.281	0.185
1100	0.019	0.294	0.063	0.16	0.234	0.236	0.212	0.276	0.187
1300	0.018	0.29	0.064	0.158	0.231	0.247	0.215	0.272	0.187
1500	0.018	0.287	0.064	0.156	0.23	0.253	0.216	0.27	0.187
1700	0.018	0.285	0.064	0.154	0.23	0.257	0.217	0.269	0.187
1900	0.018	0.283	0.064	0.153	0.23	0.258	0.217	0.269	0.186
2100	0.018	0.282	0.064	0.153	0.23	0.259	0.217	0.268	0.186
Phi4									
500	0.017	0.286	0.065	0.165	0.237	0.169	0.225	0.285	0.181
700	0.019	0.297	0.066	0.166	0.243	0.198	0.219	0.285	0.187
900	0.018	0.298	0.065	0.164	0.244	0.218	0.222	0.291	0.19
1100	0.018	0.3	0.067	0.163	0.249	0.236	0.223	0.295	0.194
1300	0.017	0.301	0.068	0.162	**0.251**	0.251	0.227	0.299	0.197
1500	0.016	0.296	0.067	0.158	0.246	0.261	0.225	0.292	0.195
1700	0.015	0.288	0.066	0.154	0.24	0.267	0.225	0.285	0.193
1900	0.014	0.281	0.065	0.151	0.238	0.27	0.227	0.28	0.191
2100	0.014	0.276	0.065	0.148	0.235	**0.271**	0.229	0.276	0.189

Table 18: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the full prompt with instruction 4. Bold values indicate the best score within this table for each metric.

#n	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Meteor	AlignScore	MEDCON	Overall
Gemma2									
500	0.018	0.305	0.071	0.177	0.258	0.179	0.281	0.315	0.2
700	0.021	0.324	0.075	**0.183**	0.261	0.211	0.271	0.315	0.208
900	0.021	**0.327**	**0.076**	0.182	0.258	0.228	0.26	0.312	0.208
1100	0.021	0.323	**0.076**	0.179	0.252	0.236	0.255	0.305	0.206
1300	0.02	0.317	0.075	0.175	0.248	0.238	0.252	0.3	0.203
1500	0.02	0.313	0.074	0.173	0.246	0.238	0.251	0.299	0.202
1700	0.02	0.311	0.074	0.171	0.246	0.239	0.25	0.298	0.201
1900	0.019	0.31	0.074	0.17	0.246	0.24	0.25	0.297	0.201
2100	0.019	0.309	0.074	0.17	0.246	0.24	0.251	0.297	0.201
Gemma3									
500	0.009	0.289	0.063	0.167	0.25	0.171	0.283	0.293	0.191
700	0.013	0.309	0.067	0.173	0.259	0.204	**0.285**	0.312	0.203
900	0.014	0.321	0.071	0.178	**0.264**	0.23	**0.285**	**0.318**	0.21
1100	0.015	0.322	0.072	0.176	0.262	0.244	0.278	0.316	**0.21**
1300	0.015	0.317	0.071	0.172	0.257	0.248	0.269	0.311	0.207
1500	0.014	0.31	0.069	0.167	0.25	0.248	0.265	0.304	0.203
1700	0.014	0.305	0.068	0.164	0.247	0.247	0.262	0.299	0.201
1900	0.014	0.301	0.067	0.162	0.245	0.247	0.261	0.297	0.199
2100	0.014	0.299	0.067	0.161	0.245	0.247	0.261	0.296	0.199
LLAMA3									
500	0.019	0.293	0.066	0.171	0.252	0.174	0.25	0.277	0.188
700	0.021	0.306	0.067	0.174	0.258	0.205	0.242	0.287	0.195
900	**0.022**	0.307	0.068	0.171	0.254	0.226	0.236	0.29	0.197
1100	0.021	0.302	0.068	0.167	0.248	0.241	0.237	0.286	0.196
1300	0.02	0.294	0.067	0.163	0.243	0.251	0.242	0.283	0.195
1500	0.019	0.286	0.067	0.158	0.239	0.258	0.246	0.277	0.194
1700	0.018	0.277	0.065	0.154	0.236	0.261	0.249	0.274	0.192
1900	0.017	0.268	0.063	0.15	0.234	0.261	0.253	0.272	0.19
2100	0.016	0.261	0.062	0.146	0.232	0.261	0.255	0.27	0.188
Mistral									
500	0.013	0.276	0.056	0.158	0.234	0.171	0.228	0.265	0.175
700	0.015	0.287	0.056	0.158	0.233	0.197	0.224	0.273	0.18
900	0.015	0.289	0.057	0.156	0.23	0.217	0.222	0.276	0.183
1100	0.016	0.289	0.059	0.155	0.227	0.234	0.222	0.28	0.185
1300	0.016	0.287	0.059	0.154	0.226	0.244	0.222	0.281	0.186
1500	0.015	0.285	0.06	0.153	0.227	0.251	0.221	0.279	0.186
1700	0.016	0.283	0.06	0.152	0.228	0.255	0.222	0.279	0.187
1900	0.016	0.282	0.061	0.152	0.228	0.256	0.222	0.278	0.187
2100	0.016	0.282	0.061	0.152	0.228	0.257	0.222	0.279	0.187
Phi4									
500	0.015	0.284	0.062	0.164	0.241	0.167	0.284	0.26	0.185
700	0.017	0.299	0.062	0.166	0.249	0.195	0.274	0.282	0.193
900	0.019	0.306	0.065	0.167	0.253	0.22	0.268	0.29	0.199
1100	0.018	0.306	0.064	0.164	0.253	0.237	0.261	0.295	0.2
1300	0.017	0.302	0.064	0.159	0.247	0.249	0.252	0.293	0.198
1500	0.015	0.294	0.062	0.154	0.239	0.254	0.25	0.286	0.194
1700	0.015	0.286	0.062	0.15	0.234	0.259	0.251	0.281	0.192
1900	0.014	0.28	0.062	0.147	0.232	0.261	0.253	0.277	0.191
2100	0.014	0.275	0.061	0.145	0.231	**0.262**	0.254	0.274	0.189

Table 19: Variation in evaluation scores of fine-tuned model–generated outputs based on the first n characters using the truncated prompt 4. Bold values indicate the best score within this table for each metric.

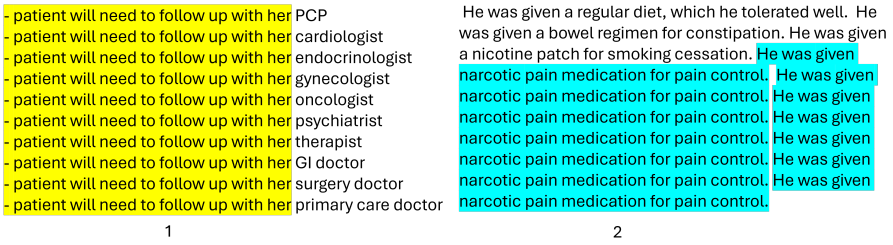


Figure 1: 1. Example of Structural Repetition, 2. Example of Content Repetition

sets. Similarly, Tables 12–18 report the scores for zero-shot prompting with full input across the same four instruction sets.

7. Different types of repetition and CUS plot

Figure 1 illustrates two distinct types of repetition observed in the generated summaries: structured repetition and content repetition. In the case of structured repetition, shown in panel (1), a structured field is repeated with differing values, highlighted in yellow. Panel (2) demonstrates content repetition, where an entire phrase is repeated verbatim, highlighted in blue.

Figure 2 presents the Comprehensive Uniqueness Score (CUS) results. The plots correspond to CUS scores for zero-shot prompting with both full and truncated input, using instruction sets 1, 2, and 3.

In Figure 3 we compute, segment-wise n-gram diversity in gold-standard discharge summaries from the training set. The figure analyzes lexical diversity over overlapping token segments (0–500, 200–700, . . . , 1600–2100) of the combined BHC and DI text from the gold-standard target summaries used for training. Unique n-gram density is computed as the ratio of distinct to total n-grams, with separate curves shown for n-gram orders n = 2–10 to illustrate variation across document progression. In addition, a composite n-gram diversity measure is reported by multiplicatively combining unique n-gram densities across n-gram orders, both including all n-grams (n = 2–10) and excluding bigrams (n = 3–10) to emphasize higher-order structural repetition. Dashed lines indicate linear trends across segments.

8. SummaC scores

The SummaC results are reported in Tables 20, 21, and 22, corresponding to the fine-tuning, zero-shot with full input, and zero-shot with truncated input settings, respectively.

9. CO2 emission related to finetuning

Experiments were conducted on NVIDIA A40 GPU infrastructure hosted in Australia. We estimated the carbon emissions associated with fine-tuning four models on the training data for both tasks, BHC and DI generation. Table 23 reports the resulting carbon emissions in kilograms of CO_2 equivalent.

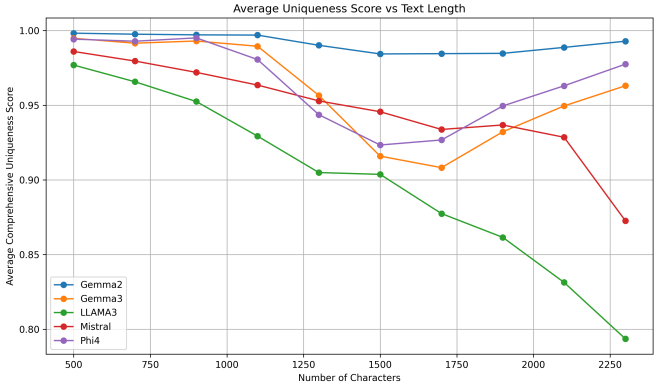
Emission estimates were obtained using the CodeCarbon emissions tracker [10, 11], an open-source framework for tracking and reporting the carbon footprint of machine learning experiments.

References

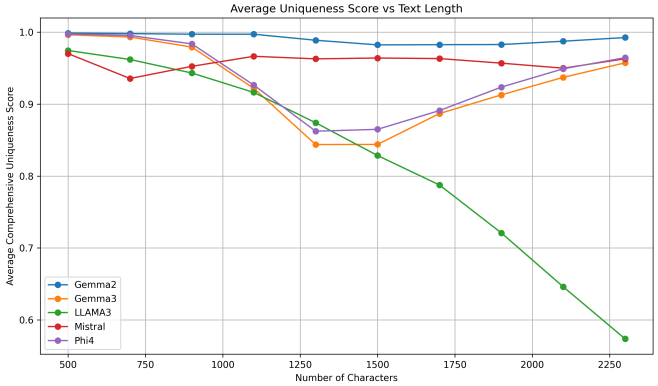
[1] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-W. H. Lehman, L. A. Celi, R. G. Mark, MIMIC-IV, a freely accessible electronic health record dataset, Scientific Data 10 (1) (2023) 1. doi:10.1038/s41597-022-01899-x.

[2] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) E215–220. doi:10.1161/01.cir.101.23.e215.

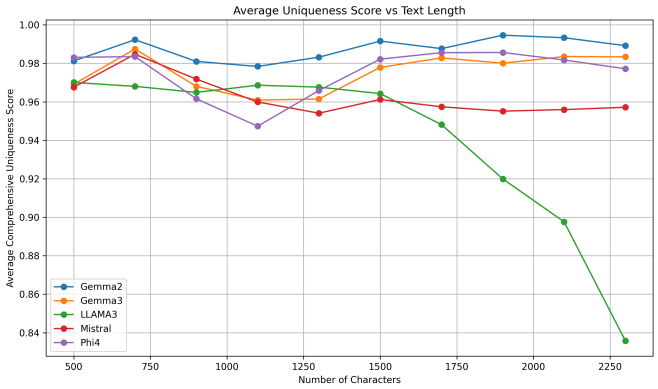
[3] J. Xu, Z. Chen, A. Johnston, L. Blankemeier, M. Varma, J. Hom, W. J. Collins, A. Modi, R. Lloyd, B. Hopkins, C. Langlotz, J.-B. Delbrouck, Overview of the First Shared Task on Clinical Text Generation: RRG24 and "Discharge Me!" (2024) 85–98doi:10.18653/v1/2024.bionlp-1.7.
URL https://aclanthology.org/2024.bionlp-1.7/



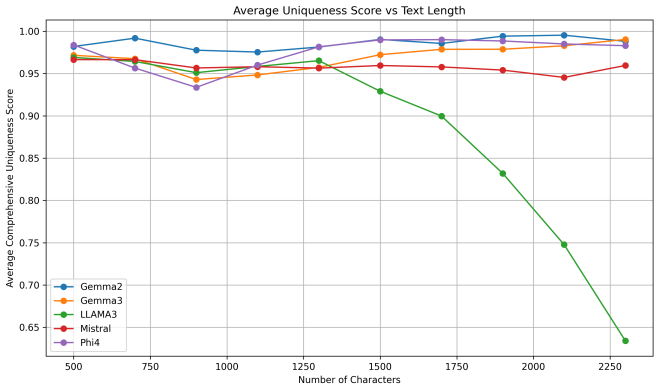
(a) CUS graph for zero shot prompt generated summary with full prompt, with instruction 1



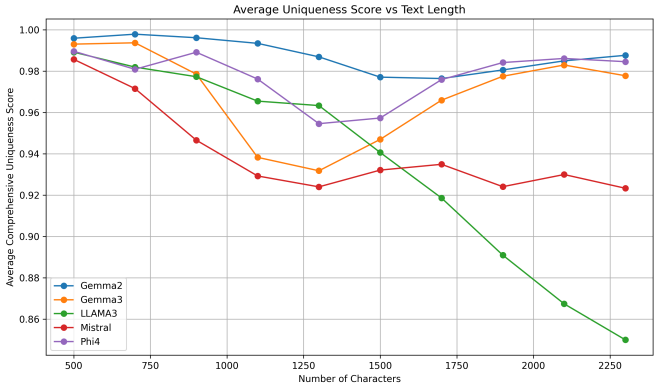
(b) CUS graph for zero shot prompt generated summary with truncated prompt, with instruction 1



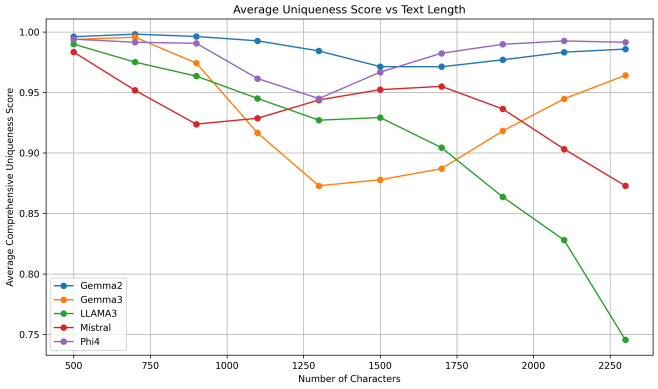
(c) CUS graph for zero shot prompt generated summary with full prompt, with instruction 2



(d) CUS graph for zero shot prompt generated summary with truncated prompt, with instruction 2



(e) CUS graph for zero shot prompt generated summary with full prompt, with instruction 3



(f) CUS graph for zero shot prompt generated summary with truncated prompt, with instruction 3

Figure 2: Comprehensive uniqueness score (CUS) computed on 500 character chunk as stated in

	SummaC-ZS				SummaC-Conv				SummaC-ZS				SummaC-Conv			
	0-500	**500-1k**	**1k-1.5k**	**1.5k-end**	**0-500**	**500-1k**	**1k-1.5k**	**1.5k-end**	**0-500**	**0-1k**	**0-1.5k**	**full**	**0-500**	**0-1k**	**0-1.5k**	**full**
Fine tune - full input																
Gemma2	**-0.413**	-0.474	-0.477	-0.446	0.485	**0.486**	0.47	0.481	**-0.413**	-0.45	-0.464	-0.473	**0.485**	0.457	0.455	0.456
Llama3	**-0.445**	-0.495	-0.506	-0.51	0.471	**0.483**	0.469	0.479	**-0.445**	-0.479	-0.494	-0.505	**0.471**	0.45	0.448	0.448
Mistral	**-0.467**	-0.502	-0.499	-0.524	0.452	**0.479**	0.478	0.461	**-0.467**	-0.49	-0.498	-0.505	**0.452**	0.437	0.436	0.437
Phi4	**-0.3**	-0.391	-0.422	-0.449	**0.704**	0.665	0.65	0.645	**-0.3**	-0.345	-0.369	-0.389	**0.704**	0.703	0.703	0.703
Fine tune - truncated input																
Gemma2	**-0.429**	-0.486	-0.498	-0.515	0.485	**0.495**	0.49	0.446	**-0.429**	-0.469	-0.488	-0.503	**0.485**	0.462	0.459	0.459
Llama3	**-0.477**	-0.519	-0.521	-0.53	0.46	0.466	**0.47**	0.468	**-0.477**	-0.509	-0.522	-0.532	**0.46**	0.437	0.433	0.434
Mistral	**-0.45**	-0.507	-0.504	-0.52	0.48	0.481	**0.483**	0.464	**-0.45**	-0.486	-0.497	-0.506	**0.48**	0.46	0.457	0.458
Phi4	**-0.399**	-0.416	-0.434	-0.47	0.495	**0.509**	0.504	0.487	**-0.399**	-0.417	-0.433	-0.463	**0.495**	0.478	0.476	0.476

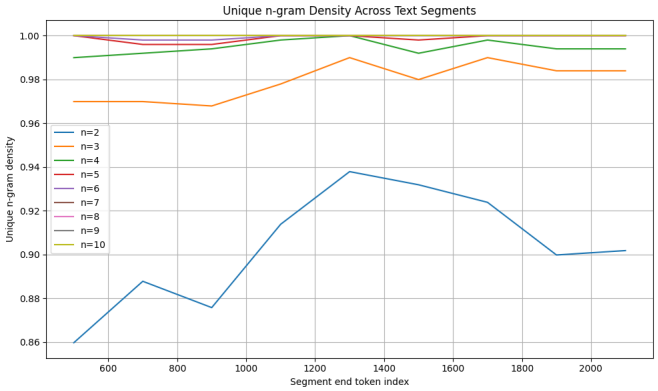
Table 20: SummaC scores for different finetuned models, computed over character-based chunks of generated summaries. Bold numbers indicate the highest SummaC-ZS and SummaC-Conv scores among all segments for each model.

	0-500	500-1k	1k-1.5k	1.5k-end	0-500	500-1k	1k-1.5k	1.5k-end	0-500	0-1k	0-1.5k	full	0-500	0-1k	0-1.5k	full
Zero shot - full input - instruction 1																
Gemma2	-0.44	**-0.437**	-0.486	-0.533	0.396	**0.46**	0.459	0.41	**-0.44**	-0.451	-0.461	-0.474	**0.396**	0.381	0.371	0.37
Gemma3	-0.358	-0.364	**-0.337**	-0.459	0.428	0.495	**0.542**	0.473	**-0.358**	-0.37	-0.367	-0.391	0.428	**0.429**	0.421	0.421
Llama3	-0.333	-0.412	-0.368	**-0.275**	**0.572**	0.508	0.556	0.534	-0.333	-0.366	-0.369	**-0.297**	**0.572**	0.55	0.539	0.537
Mistral	-0.407	**-0.328**	-0.383	-0.427	0.419	0.512	0.507	**0.538**	-0.407	-0.349	**-0.344**	-0.349	0.419	0.446	0.45	**0.453**
Phi4	-0.425	**-0.374**	-0.407	-0.5	0.407	**0.524**	0.497	0.44	-0.425	**-0.409**	-0.412	-0.452	0.407	**0.452**	0.446	0.445
Zero shot - full input - instruction 2																
Gemma2	**-0.432**	-0.477	-0.496	-0.48	0.437	**0.484**	0.45	0.439	**-0.432**	-0.465	-0.482	-0.485	**0.437**	0.433	0.423	0.423
Gemma3	**-0.36**	-0.363	-0.414	-0.38	**0.532**	0.51	0.483	0.496	**-0.36**	-0.372	-0.391	-0.389	**0.532**	0.502	0.497	0.496
Llama3	-0.366	-0.384	-0.429	**-0.284**	**0.581**	0.59	0.515	0.527	-0.366	-0.384	-0.406	**-0.31**	0.581	**0.589**	0.579	0.579
Mistral	-0.418	-0.336	-0.348	**-0.331**	0.482	**0.583**	0.544	0.564	-0.418	-0.381	-0.363	**-0.362**	0.482	0.517	0.526	**0.531**
Phi4	**-0.367**	-0.376	-0.443	-0.403	**0.501**	0.493	0.494	0.474	**-0.367**	-0.381	-0.404	-0.412	**0.501**	0.471	0.47	0.468
Zero shot - full input - instruction 3																
Gemma2	-0.536	**-0.463**	-0.484	-0.487	0.407	**0.449**	0.426	0.418	-0.536	-0.512	-0.507	**-0.501**	**0.407**	0.384	0.374	0.374
Gemma3	-0.477	-0.471	**-0.45**	-0.467	0.382	**0.457**	0.444	0.45	**-0.477**	-0.489	-0.485	-0.487	**0.382**	0.381	0.369	0.369
Llama3	-0.357	-0.407	-0.508	**-0.306**	**0.562**	0.558	0.444	0.504	-0.357	-0.396	-0.424	**-0.332**	**0.562**	0.54	0.534	0.533
Mistral	-0.44	**-0.377**	-0.392	-0.42	0.435	**0.514**	0.509	0.487	-0.44	-0.403	-0.394	**-0.393**	0.435	0.461	0.478	**0.481**
Phi4	-0.516	**-0.458**	-0.5	-0.512	0.397	**0.482**	0.465	0.423	-0.516	**-0.503**	-0.511	-0.517	0.397	**0.404**	0.39	0.389
Zero shot - full input - instruction 4																
Gemma2	**-0.433**	-0.465	-0.492	-0.493	0.391	0.432	**0.458**	0.433	**-0.433**	-0.462	-0.472	-0.479	**0.391**	0.367	0.354	0.353
Gemma3	-0.445	-0.437	**-0.415**	-0.474	0.389	0.442	**0.483**	0.418	**-0.445**	-0.455	-0.455	-0.46	**0.389**	0.378	0.371	0.372
Llama3	-0.383	-0.36	-0.402	**-0.307**	0.546	**0.584**	0.529	0.532	-0.383	-0.381	-0.386	**-0.301**	**0.546**	0.537	0.527	0.525
Mistral	-0.424	**-0.281**	-0.325	-0.388	0.422	**0.553**	0.552	0.539	-0.424	-0.338	**-0.327**	-0.333	0.422	0.471	0.48	**0.481**
Phi4	**-0.413**	-0.422	-0.433	-0.489	0.398	0.464	**0.468**	0.44	**-0.413**	-0.437	-0.442	-0.472	0.398	**0.4**	0.384	0.383

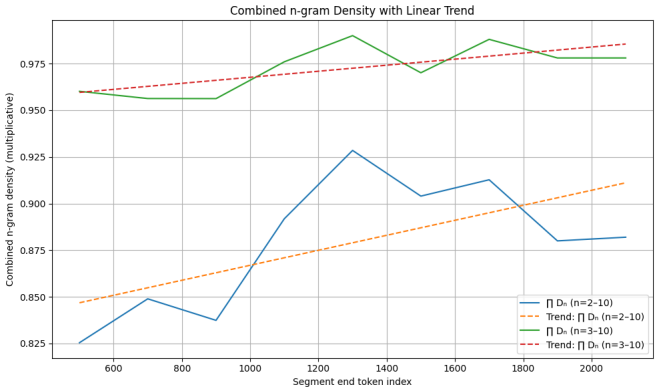
Table 21: SummaC scores for different models, computed over character-based chunks of generated summaries using zero shot approaches with full input. Bold numbers indicate the highest SummaC-ZS and SummaC-Conv scores among all segments for each model.

	SummaC-ZS				SummaC-Conv				SummaC-ZS				SummaC-Conv			
	0-500	**500-1k**	**1k-1.5k**	**1.5k-end**	**0-500**	**500-1k**	**1k-1.5k**	**1.5k-end**	**0-500**	**0-1k**	**0-1.5k**	**full**	**0-500**	**0-1k**	**0-1.5k**	**full**
Zero shot - truncated input - instruction 1																
Gemma2	**-0.371**	-0.418	-0.529	-0.611	0.445	**0.474**	0.445	0.44	**-0.371**	-0.382	-0.396	-0.41	**0.445**	0.426	0.425	0.423
Gemma3	-0.376	**-0.314**	-0.378	-0.515	0.448	**0.509**	0.487	0.453	-0.376	**-0.349**	-0.353	-0.371	**0.448**	**0.448**	0.443	0.442
Llama3	-0.387	**-0.34**	-0.385	-0.415	0.486	**0.54**	0.526	0.483	-0.387	**-0.376**	**-0.376**	-0.389	**0.486**	0.48	0.476	0.476
Mistral	-0.299	**-0.246**	-0.332	-0.357	0.524	**0.571**	0.542	0.563	-0.299	**-0.267**	-0.279	-0.289	0.524	0.549	**0.55**	0.549
Phi4	-0.442	**-0.345**	-0.421	-0.501	0.429	**0.527**	0.466	0.439	-0.442	**-0.392**	-0.411	-0.441	0.429	**0.458**	0.445	0.444
Zero shot - truncated input - instruction 2																
Gemma2	-0.433	**-0.405**	-0.41	-0.441	0.495	0.522	0.493	**0.571**	-0.433	**-0.422**	-0.425	-0.426	0.495	**0.503**	0.497	0.497
Gemma3	-0.361	**-0.342**	-0.375	-0.422	0.591	0.512	0.51	**0.52**	-0.361	**-0.358**	-0.367	-0.384	**0.591**	0.555	0.543	0.543
Llama3	-0.404	-0.396	**-0.376**	-0.38	0.518	**0.52**	0.505	0.482	-0.404	-0.397	**-0.393**	-0.41	**0.518**	0.517	0.515	0.513
Mistral	-0.442	-0.357	**-0.315**	-0.373	0.491	**0.556**	0.551	**0.556**	-0.442	-0.391	**-0.359**	-0.368	0.491	0.514	0.528	**0.531**
Phi4	**-0.381**	-0.404	-0.433	-0.46	**0.509**	0.474	0.489	0.446	-0.381	**-0.393**	-0.413	-0.436	**0.509**	0.494	0.487	0.483
Zero shot - truncated input - instruction 3																
Gemma2	-0.504	**-0.434**	-0.538	-0.505	0.426	**0.428**	0.419	0.402	-0.504	-0.487	-0.491	**-0.483**	**0.426**	0.388	0.382	0.382
Gemma3	-0.511	-0.431	**-0.419**	-0.446	0.387	**0.47**	0.45	0.44	-0.511	-0.486	**-0.47**	-0.471	0.387	**0.39**	0.379	0.378
Llama3	-0.476	**-0.397**	-0.47	-0.474	0.441	**0.502**	0.458	0.429	-0.476	**-0.446**	-0.458	-0.465	**0.441**	0.436	0.423	0.422
Mistral	-0.401	**-0.37**	-0.448	-0.458	0.459	**0.513**	0.479	0.456	-0.401	**-0.383**	-0.394	-0.4	0.459	**0.479**	0.477	0.477
Phi4	-0.514	**-0.448**	-0.492	-0.498	0.402	**0.473**	0.455	0.413	-0.514	**-0.496**	-0.5	-0.505	**0.402**	0.399	0.388	0.387
Zero shot - truncated input - instruction 4																
Gemma2	**-0.451**	-0.471	-0.5	-0.52	0.408	**0.434**	0.43	0.431	**-0.451**	-0.465	-0.473	-0.477	**0.408**	0.381	0.377	0.375
Gemma3	-0.459	**-0.401**	-0.409	-0.503	0.409	**0.477**	**0.477**	0.433	-0.459	-0.447	**-0.442**	-0.451	**0.409**	0.401	0.388	0.388
Llama3	-0.36	**-0.32**	-0.426	-0.449	0.439	**0.495**	0.461	0.454	-0.36	**-0.346**	-0.378	-0.398	0.439	**0.44**	0.426	0.424
Mistral	-0.374	**-0.293**	-0.37	-0.443	0.474	**0.552**	0.504	0.502	-0.374	**-0.331**	-0.344	-0.353	0.474	**0.508**	0.498	0.495
Phi4	-0.415	**-0.374**	-0.428	-0.516	0.421	**0.462**	0.461	0.419	-0.415	**-0.411**	-0.422	-0.464	**0.421**	0.398	0.387	0.385

Table 22: SummaC scores for different models, computed over character-based chunks of generated summaries using zero shot approaches with truncated input.



(a) Unique n-gram density across overlapping document segments. The figure shows unique n-gram density (|unique n-grams| / |total n-grams|) computed over overlapping token segments (0–500, 200–700, . . . , 1600–2100) of the combined BHC and DI text. Separate curves correspond to n-gram orders n = 2–10. The x-axis denotes the segment end index, and the y-axis denotes the n-gram density for the corresponding segment.



(b) Multiplicative composite n-gram density over document progression. For each overlapping token segment, unique n-gram density is computed as |unique n-grams| / |total n-grams| and combined multiplicatively across n-gram orders. Results are shown for n = 2–10 and for n = 3–10 (excluding bigrams). Segment end indices define the x-axis. Dashed lines represent least-squares linear trends.

Figure 3: Uniqueness density computed on target summary sections (BHC and DI) of training set.

		Gemma2	Llama3	Mistral	Phi4
CO_2 **in kg**	**Truncated Input**	10.5	5.53	6.46	10.01
	Full Input	17.89	8.8	11.39	17.49

Table 23: Comparison of carbon emissions (in kg) during fine-tuning of four language models with full input and truncated input settings.

[4] R. Flesch, How to write plain english, University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.[Retrieved 5 February 2016] (1979).

[5] H. Damm, T. M. G. Pakull, B. Eryılmaz, H. Becker, A. Idrissi-Yaghir, H. Schäfer, S. Schultenkämper, C. M. Friedrich, WisPerMed at "Discharge Me!": Advancing Text Generation in Healthcare with Large Language Models, Dynamic Expert Selection, and Priming Techniques on MIMIC-IV (2024) 105–121doi:10.18653/v1/2024.bionlp-1.9.
URL https://aclanthology.org/2024.bionlp-1.9/

[6] V. Socrates, T. Huang, X. Ai, S. Fereydooni, Q. Chen, R. A. Taylor, D. Chartash, Yale at "Discharge Me!": Evaluating Constrained Generation of Discharge Summaries with Unstructured and Structured Information, in: D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 724–730. doi:10.18653/v1/2024.bionlp-1.64.
URL https://aclanthology.org/2024.bionlp-1.64/

[7] J. Liu, A. Nicolson, J. Dowling, B. Koopman, A. Nguyen, e-Health CSIRO at "Discharge Me!" 2024: Generating Discharge Summary Sections with Fine-tuned Language Models (2024) 675–684doi:10.18653/v1/2024.bionlp-1.59.
URL https://aclanthology.org/2024.bionlp-1.59/

[8] H. Wu, P. Boulenger, A. Faure, B. Céspedes, F. Boukil, N. Morel, Z. Chen, A. Bosselut, EPFL-MAKE at "Discharge Me!": An LLM System for Automatically Generating Discharge Summaries of Clinical Electronic Health Record, in: D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 696–711. doi:10.18653/v1/2024.bionlp-1.61.
URL https://aclanthology.org/2024.bionlp-1.61/

[9] M. Lyu, C. Peng, D. Paredes, Z. Chen, A. Chen, J. Bian, Y. Wu, UF-HOBI at "Discharge Me!": A Hybrid Solution for Discharge Summary Generation Through Prompt-based Tuning of GatorTronGPT Models (2024) 685–695doi:10.18653/v1/2024.bionlp-1.60.
URL https://aclanthology.org/2024.bionlp-1.60/

[10] A. Lacoste, A. Luccioni, V. Schmidt, T. Dandres, Quantifying the carbon emissions of machine learning, Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019 (2019).

[11] K. Lottick, S. Susai, S. A. Friedler, J. P. Wilson, Energy usage reports: Environmental awareness as part of algorithmic accountability, Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019 (2019).